Before the Federal Communications Commission
Washington, D.C. 20554

In the Matter of	)	
	)	
Petition for Declaratory Ruling and Petition	)	RM8503
for Rulemaking on Live Closed Captioning	)	CG Docket 05-231
Quality Metrics and the Use of Automated	)	
Speech Recognition Technologies	)	

VITAC, a full-service captioning company with more than 33 years of experience in the captioning industry, strongly endorses and supports the Petition for Declaratory Ruling and/or Rulemaking on Live Closed Captioning Quality Metrics and the Use of Automatic Speech Recognition Technologies before the Federal Communications Commission that asks the Federal Communications Commission to begin developing objective, technology- and methodology-neutral metrics for live captioning quality, and issue a declaratory ruling on the use of ASR technologies.

Captions play a crucial function in the daily lives of more than 50 million deaf and hard-of-hearing Americans[1], and there is little doubt that clear, concise, accurate captions are essential for those who need them the most. Whether across traditional broadcast or through streaming video or social media platforms, captions are critical in ensuring accessible services to entertainment, education, employment, news, threats to human life and health, and emergency information.

Like other captioning and accessibility providers upon whom the deaf and hard-of-hearing community rely every day, VITAC has conducted extensive research into the quality of captions

[1] Frank Lin (2011, November). Hearing Loss Prevalence in the United States. Retrieved from https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/1106004.

created by automatic speech recognition (ASR) systems. It is our belief that there needs to be two separate and distinct definitions utilized when ASR is employed.

The first definition for usage of an ASR engine is "Supervised ASR," wherein humans who are highly trained and skilled professionals are employed to interact with an ASR engine that is trained to recognize and accurately caption the applicable spoken content.

The second usage instance definition is "Unsupervised ASR," where an ASR engine is employed to create captions in a work environment without human intervention, as is being tried without a great deal of accuracy or success in several broadcast sites today. VITAC provides over 3,400 hours of Supervised ASR every week, and we have been utilizing ASR successfully in commercial broadcast sites of all sizes and geographies for more than a decade. We are experienced users of Supervised ASR, and we are highly qualified to address the relative benefits and shortcomings of both Supervised and Unsupervised ASR usage.

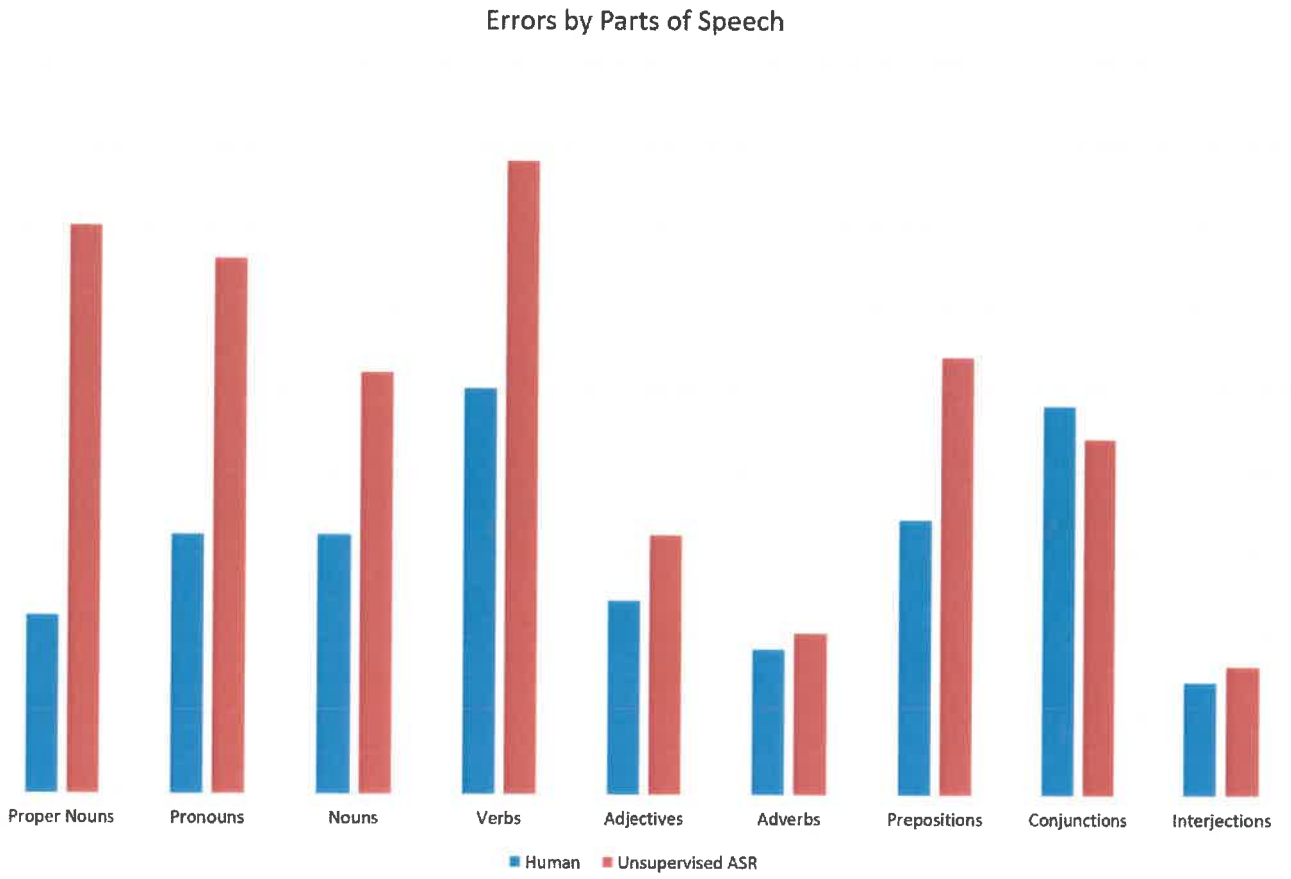
During the course of our successful commercial use of ASR technology, we've determined that five-minute video segments of approximately 1,000 words provide a perfect sample in determining caption accuracy, no matter the tools, techniques, and technology employed to generate captions.

Our research and analysis conclusively demonstrate that live captions, those resulting from a live broadcast event, created by Unsupervised ASR routinely fail to meet expectations or Caption Quality Best Practices as outlined by the FCC (Report and Order in CG Docket No. 05-231, FCC 14-12, section 79.1(k)(2),(3) and (4).)

Unsupervised ASR routinely misses proper punctuation, speaker changes, sound effects, music, and lyrics. That is why human captioners "supervise and instruct" ASR on matters of punctuation, speaker changes, sound effects, and other nuances of human dialogue.

Although both human and Unsupervised ASR-generated captions contain missing words, not all missing words are equal. Our research consistently proves that Unsupervised ASR misses more important words compared to those missed by human captioners. Unsupervised ASR most frequently misses proper nouns, pronouns, nouns, and verbs. Human captioners also omit these same parts of speech, although much less frequently.

We have measured, reviewed, and analyzed a number of Unsupervised ASR-captioned news segments, and scored the results for such things as wrong words, missing words, and missing speaker IDs, but also looked for what parts of speech were missed. The chart below shows the Unsupervised ASR vs. human captioner results.



If the fundamental purpose of captions is to convey meaning, then words that carry the most meaning cannot be omitted. Therefore, captions must be well constructed and presented in a way that makes consuming and understanding them by the viewer easy.

One of the biggest challenges is in presenting captions at a speed that ensures that all viewers have enough time to read and comprehend them. A 2019 study[2] by Marc Brysbaert of Ghent University found that the average silent reading speed for most adults in English is around 235

[2] Marc Brysbaert (2019, April). How Many Words Do We Read Per Minute? A Review and Meta-Analysis of Reading Rate. Retrieved from https://psyarxiv.com/xynwg/.

words per minute. Reading is a complex process that involves a variety of factors, but caption viewers face an additional challenge when processing text – namely the need to read and understand captions while also assimilating the interactive content or depictions simultaneously displayed.

If caption consumers are served text that too quickly appears on the screen, it hinders comprehension and usefulness. Likewise, if captions, such as those created by Unsupervised ASR, contain missing words, wrong words, and lack punctuation, viewers will lose precious time trying to decipher the caption and miss the message altogether. While this definitely is a distraction in the context of enjoying content that is entertaining, it could lead to a serious, potentially fatal outcome at times of emergency or crisis.

Unsupervised ASR tries to caption everything on the screen, guessing at words and largely failing. This often results in portions of dialogue, including critical and non-critical dialogue, being dropped in order for the machine to catch up to the current discussion. The end result is that viewers are uninformed about the information being conveyed.

A human captioner may make a conscious decision to drop an inconsequential word (perhaps a definitive article) from a sentence to better deliver captions that flow with realtime dialogue and, in the process, provide a more accurate reflection of a sentence's meaning for caption consumers. The resulting captions provided by humans, both natively and using Supervised ASR, reflect a conscious decision to edit complex, unclear realtime text, while maintaining the meaning, all to the objective of clearer communications and better caption understanding.

Though it's desirable and advantageous to employ new technologies to advance the art and delivery of captions to those who rely upon them for inclusion in every aspect of their lives, it is incumbent upon us to do so wisely, and ensure that those technologies are a true benefit to those who depend upon the timely, accurate, and dependable use of any technology.

Unsupervised ASR captions are not a proven solution at this time. They are not accurate to the point of reliability and, therefore, incapable of meeting the critical accessibility needs of the community who relies on them.

We urge the FCC to consider and take action on the Petition for Declaratory Ruling and/or Rulemaking on Live Closed Captioning Quality Metrics and the Use of Automatic Speech Recognition Technologies as filed, and set uniform quality standards to meet the needs of the deaf and hard-of-hearing, eliminating any further challenges experienced by those who regularly and essentially rely upon captions, whether they be created natively by humans, humans in concert with machines, or machines alone.

Respectfully Submitted,

VITAC
8300 E Maplewood Avenue, Suite 310
Greenwood Village, CO 80109

P. Kevin Kilroy

P. Kevin Kilroy, CEO

October 7, 2019